PCT/GB2003/001372

#2

The Patent Office

INVESTOR IN PEOPLE

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

REC'D 22 MAY 2003
WIPO PCT

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed Andrew George

Dated 30 April 2003

Patents Form 1/77

Patents Act 1977
(Rule 16)

The Patent Office

1/77

THE PATENT OFFICE
L
28 MAR 2002
LONDON

28 MAR 2002

The Patent Office

Cardiff Road
Newport
South Wales
NP9 1RH

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

1.	Your reference	P/63624.GBA/CAMLAB
2.	Patent application number *(The Patent Office will fill in this part)*	0207507.5 ~~10MAR02 E707632-3 D00314~~ P01/7700 0.00-0207507.5
3.	Full name, address and postcode of the or of each applicant *(underline all surnames)*	Marconi Corporation plc One Bruton Street ~~London W1J 6AQ~~
	Patents ADP number *(if you know it)*	797959402
	If the applicant is a corporate body, give the Country/state of its incorporation	England
4.	Title of the invention	An apparatus for providing communications network resource
5.	Name of your agent *(if you have one)*	Nigel George McGowan
	"Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*	Marconi Intellectual Property, Marrable House, The Vineyards, Gt. Baddow, Chelmsford Essex CM2 7QS
	Patents ADP number *(if you know it)*	7873300001

		Country	Priority application number *(if you know)*	Date of filing *(day / month / year)*
6.	If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number			

		Number of earlier application	Date of filing *(day / month / year)*
7.	If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application		

8.	Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:* *a) any applicant named in part 3 is not an inventor, or* *b) there is an inventor who is not named as an applicant, or* *c) any named applicant is a corporate body.* *See note (d))*	Yes

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form	
Description	14
Claim(*s*)	1
Abstract	1 DMC
Drawings	3 +3

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	1
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	

11. I/We request the grant of a patent on the basis of this application.

Signature N.G.McGowan Date 28th March 2002

12. Name and daytime telephone number of person to contact in the United Kingdom 01245 707602

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

An apparatus for providing communications network resource

is invention relates to an apparatus for providing communications network source.

the past, networks -- in particular those used to support the Internet -- uld share resources: the buffers of the routers and the line capacity of the nnections between routers and hosts, between all network users. In the modern twork it is desirable to divide the resource between different network traffic pes. Rather than the shared service of the common Internet, network providers sh to divide the resource between the traffic types based on other aracteristics such as their willingness to pay for service, their need for ffering service quality or some combination of the two.

twork-elements (routers and switches) that employ resource partitioning, such the division of link bandwidth between different classes of traffic, have ed a scheduler that fixed, as part of its algorithm, the amount of the source (outgoing bandwidth) to be allocated to each class, e.g. [1-3].

a result, traffic queued in network routers that was not serviced could be ther delayed or lost as the queue filled and overflowed. In such a scheme the ~~sources of buffer space and the service weights of the scheduler were~~ located according to policy e.g. based on a simple priority scheme or with an signed weighting based on the value of each traffic class.

st interest in traffic characterisation (e.g. with respect to [4-6]) has noted e difficulties inherent in this approach. These schemes are difficult both to nfigure initially and to keep correct under changing network demand - as a sult such schemes waste resource and are limited in the complexity of services fered.

e present invention proposes the use of a Measurement-Based Estimator (MBE) of ndwidth requirements to enhance the performance of resource scheduling. The tivation of this approach was that it could be retro-fitted to current twork-elements without significant drawback. Dynamic allocation could be added any network-element as needs dictated. In such a per-network-element proach, the MBE is used to compute the precise resource weighting of the mand of each traffic type. This estimate is then used to adjust the weights of neduling algorithms as well as adjust the depth and behaviour of buffering.

e main contributions are twofold. Firstly, building upon the substantial work at has been invested in Measurement-based Admission Control (MBAC) (e.g. [7-]), this work adapts an MBAC Algorithm to continuously provide measurement-sed estimations of bandwidth requirements. The second major contribution is to aluate a dynamic resource allocator built upon this MBE. Importantly, the aluation is performed using an actual implementation in an active test vironment.

block diagram of the system is given in Figure 1. The dynamic allocation stem consists of a method for computing the demand of each traffic class. In e implementation presented here, the Measurement-Based Estimator computes the mand of each class using measurements of line utilisation, these estimates of mand are then provided to the dynamic resource allocator. The dynamic resource locator is able to configure the weights of a scheduler, (e.g. the weights of weighted-round-robin scheduler), and the maximum buffer depth that a rticular traffic class may use. The network-element will impose these nfigured restrictions upon the classes of traffic as they are multiplexed onto e outgoing link.

e remainder of this description is structured as follows. Section 2 discusses fferentiated services with particular reference to two differentiated service

models which are introduced there. Section 2 also discusses allocation policy. Section 3 then discusses the relevant buffer and scheduler technologies as well as discussing the measurement-based estimator used in construction of the implementation.

The details of the implementation are given in Section 4, while the test environment and experimental method are given in Section 5. Section 6 details the results of experiments illustrating the operation of the dynamic resource allocator and Section 7 provides concluding remarks as well as noting several areas for future improvement.

2 A Background to Differentiated Service

Differentiated services may simply be defined as he ability of a network to offer two or more types of network behaviour to the network users. Examples of such networks include a network offering low-latency or a network that offers low-loss. The combined approach of admission control and an appropriate scheduling algorithm has long been considered central to supplying Quality-of-Service (QoS) in an integrated services network [13]. However, admission control is not generally considered practical in networks such as the modern Internet. Attempts at introducing Admission Control techniques (e.g. IntServ/ RSVP [14]) are considered largely impractical to implement on a wide scale.

Networks wishing to provide QoS but without explicit admission control are a central idea of the approach of the differentiated-services network architecture, DiffServ [15,16]. Flows carried in a differentiated services system such as DiffServ do not receive an individual guarantee of resources. Instead, a guarantee is made to the class of traffic to which each flow belongs. The class of traffic will receive all the resources it requires but individual flow properties and flow interaction will mean that the per-flow resourcing will be only statistical in nature. This means that at any instant one particular flow may receive greater or fewer resources than it requires.

The following concentrates upon a single network-element, e.g. a core router, carrying several, pre-classified, classes of traffic. Two particular types of service differentiation are used as examples and these two schemes are now explained in detail.

2.1 Olympic Service Differentiation

In the Olympic Service implemented using Assured Forwarding described in [17], there exists three classes: bronze, silver, and gold. Traffic in each of these three classes is configured so that the gold class experiences lighter load than the silver and the silver experiences lighter load than the bronze. While for bandwidth allocation a simple implementation may assign a fixed quantity of resource to each class, perhaps 50% to Gold, 30% to silver and 20% to bronze. Such a fixed allocation will neglect the actual requirements of each traffic class.

The benefit of the dynamic allocation scheme of the present invention is that it is able to ensure Gold demands are met in priority to Silver demands and in-turn meeting Silver demands while the remainder service is given to Bronze. Yet such a scheme adapts to the current requirements of each class. Thus if Silver is not using its total allocation, this left-over is made available to Bronze. Such a scheme permits minimal waste of resource while allowing the construction of new services such as a BE class that receives resource only when the higher priority gold, silver and bronze classes have received their required allocations.

Section 6.1 illustrates the operation of the implemented measurement-based allocator providing an olympic service.

Orthogonal Service Differentiation

contrast to a set of classes each having a different level of requirement in .same resource-type, an alternative set of offerings may be the combination traffic classes, each with differing requirements of different resources. An mple of such orthogonal combinations of service would be a low-loss service . a low-delay (or low delay-variation) service. This pair of traffic classes a combination of the assured and expedited forwarding classes [17,18] of fServ.

tion 6.2 gives results of the implemented measurement-based allocator viding an orthogonal set of services: a low-loss service, a delay-bound vice and a best-effort service with access to the remaining network resource.

Allocation Policy

use of a programmable, dynamic network-element able to adapt to the changing uirements of network traffic allows considerable scope for a sophisticated icy allocating the available resources to traffic requirements. Two examples broad allocation policy are the Olympic and orthogonal service ferentiation given in Sections 2.1 and 2.2, however a policy must be more plete.

resolution procedure when network resources are over (or under) allocated t be specified in the allocation policy. One example of such a resolution cedure would be through the use of a priority mechanism. In such a scheme, top priority traffic class must be satisfied completely and then the next hest priority and so on. In the Olympic service differentiation it is clear t Gold will take precedence over all below it, Silver will take precedence r all but Gold, and so on. As an alternative to priority ordering, a second mple for under-resourcing would be to diminish the actual resource to all sses of traffic, thus the drawback of under-resourcing is shared portionally among the competing classes.

the case of over-resourcing, where more resource is available than required, solution adopted may be to share the excess bandwidth evenly among the ferent traffic classes. An alternative approach for over-resourcing may be to ocate the excess resource to a best effort class of traffic, one that would y receive resource when all other classes had received their allocation.

early, ample opportunity exists for complex reconciliation behaviour. For the mples of Section 6, the reconciliation behaviour is priority based for both Olympic and orthogonal differentiated service examples. In this way, the t effort service in each example is provided with service only after the mitment of resource to all other classes of traffic has been made. The ority ordering for the Olympic service is Gold, Silver, Bronze and then best ort, while the priority ordering for the orthogonal differentiated service mple is delay-constrained traffic, loss-constrained and finally the elastic ffic (using a best effort mechanism).

Approach

approach is that a network-element will use a combination of scheduler and trol to offer differentiated services. Several assumptions are made with ard to the traffic that impact how particular traffic types will be supplied ource by the network-element.

particular, a first assumption is that for delay-sensitive network traffic, kets delayed beyond the traffic's delay boundary are of no value. This lies that delay-sensitive network traffic is best-served by a combination of fer discard threshold dictated by that delay-boundary and a non-

:kconserving scheduling algorithm. In contrast, network traffic that is ınded by loss would be buffered to a depth that did not void any delay ıstraints while being served at a rate that satisfied the loss constraints. ıffic that is throughput guaranteed is considered the most trivial traffic ›e requiring only limited buffering and a fixed buffer service rate. Finally, :t-effort traffic may make use of the remaining buffer and service bandwidth ›viding a *left-over* service; in this way the best-effort may obtain :entially all network resource but without causing starvation of any resource which a guarantee has been made.

. Scheduling Algorithms

key property to allow for the several different classes of this type is a :ket scheduler with sufficient flexibility as to be able to bound the delay ' particular session incurs in addition to simply dividing up the bandwidth ;ource. The ideal scheduler is one able to emulate Generalised Processor ıring (GPS) scheduling -- one able to (infinitely) divide up resource service :ween different traffic classes; thereby bounding delay while providing ›xible service offerings [19]. The GPS algorithm is not easily implemented in ıctice however a close emulation of GPS is available to packet networks that › a fixed cell length.

› test-environment is based upon an ATM network. ATM networks use a fixed :ket length, so the use of the GPS emulating algorithm is allowed. A suitable ; emulating algorithm is Worst-case Weighted Fair Queueing Plus (WF^2Q+), ›posed in [20]. The WF^2Q+ scheduler is implemented in the network-element of › test-environment (described in [21]) providing an environment within which › dynamic allocator can be constructed.

: Buffer Management

scheduler will allow a network node to allocate link bandwidth to each ;sion. However, for services such as voice, which is delay sensitive, ıdwidth control is not enough. As noted above, flexible buffer control can ›rove the loss-rate of both packet and burst multiplexing. Therefore, buffer ıagement provides the controls over loss while also controlling packet delay.

ıtrol over the buffer capacity available to each traffic class is required if › implementation is to provide resources for loss or delay constraint as well link bandwidth.

: delay-bound sessions, packets that exceed a buffer threshold are discarded : for loss-bound or throughput-guaranteed services the arriving packets are :ked as eligible to be discarded if no further capacity remains in the total :fer pool shared among sessions. In this way the work-conserving scheduler is .e to consume resource that would otherwise be unused.

} Measurement-Based Estimator

:ther the scheduler or buffer technology per se is new, the novelty resides in › configuration of scheduler weights and buffer capacity and behaviour in ıbination with an MBE.

ideal MBE would allow three critical resource computations: firstly, a ıputation of the capacity required to maintain a given delay-bound with a ,en probability; secondly, a computation of the capacity required to maintain loss-rate, given a particular buffer size; and lastly, the buffer size ıuired to maintain a loss-rate for a given rate of service, which would be ıuired for a service with throughput guarantee. Finally, the estimator must be ıptive to changes in traffic and flexible to changes in the traffic classes.

Estimators such as that proposed in [22] initially seem ideal for the task because they are able to combine a series of measurements with any two of the input parameters of buffer size, loss rate, or effective bandwidth and compute an estimate of the third parameter. However, the estimator of [22] depends critically upon a traffic-dependent tuning value and no robust mechanism currently exists for computing this value.

Simpler measurement-based estimators such as [9,8] or any estimator based upon a bufferless model of the network requires the computation of a complicated surface relating the desired outcome to the tuning parameter for each traffic-type. Additionally, this surface would need to exist in multiple dimensions in order to account for changes in each of link bandwidth, loss rate and queue size.

To allow ease-of-use, the MBE of the dynamic allocator must offer some relationship between calibrated controls(e.g. servicerate, loss-rate and buffer-size) and the traffic behaviour to be of use. Of equal importance is that the MBE must allow for the statistical nature of the measurement while being implementable with realistic demands on memory and processing as well as realistic demands on the measurements themselves. Accounting for each of these aspects lead to the selection of the algorithm of Knightly and Qiu [10,24] as appropriate.

Proposed originally as an admission control algorithm, the estimation component has been extracted from the admission control framework. A precis of the original algorithm is given here.

The *traffic envelope* approach embraces the central issue that to characterise the rate of a particular traffic flow a period must be specified over which that characterisation is conducted. As a result this MBE, is able to characterise traffic over a series of time periods. The intention of this multi-period characterisation is to represent the short-term burstiness of traffic as well as that of the longer-term variation of the aggregate due to measurement error and longer time-scale fluctuations.

Firstly it is assumed that there exists a basic measurement period, τ -- possibly imposed by physical measurement limitations. Measurements may be taken over a multiple of this period and thus $I_{1,2,\ldots,T} = 1,2,\ldots,T \times \tau$. Thus, if the traffic activity on a link over the interval $[s, s+I_k]$ is represented as $X[s, s+I_k]$ then $\frac{X[s, s+I_k]}{I_k}$ is the rate over that particular period. [10] noted that the peak rate over any interval of length I_k can be given by $R_k = \max_s X[s, s+I_k]$. This allows the specification of the *maximal rate envelope*: a set of rates R_k that represent the maximum rate of the flow for each of the intervals I_k.

The activity in time slot t is represented as x_t such that $x_t = X[t\tau, (t+1)\tau]$. This allows a definition of the maximal rate envelope for the past T time slots from the current time t as

$$R_k^1 = \frac{1}{k\tau} \max_{t-T+k \le s \le t} \sum_{u=s-k+1}^{s} x_u \tag{1}$$

for $k = 1,2,\ldots,T$. The envelope $R_k^1, k = 1,\ldots,T$ describes the aggregate maximal rate envelope over intervals of length $I_k = k\tau$ in the most recent $T \cdot \tau$ seconds. [10]

assert that this will describe short time-scale burstiness along with autocorrelation structure present in the flow.

If every $T \cdot \tau$ periods the current envelope is updated $R_k^n \leftarrow R_k^{(n-1)}$ for $k = 1,2,\ldots,T$ and $n = 2,\ldots,N$, then a new envelope R_k^1 is computed using Equation 1. This allows the empirical mean $\overline{R_k}$ of the R_k^m's to be computed as $\Sigma_{m=1}^{M} \frac{R_k^m}{M}$. In turn this allows the variance between envelopes for the past M windows of time $T \cdot \tau$ to be computed using

$$\sigma_k^2 = \frac{1}{M-1} \sum_{m=1}^{M} \left(R_k^m - \overline{R_k} \right)^2 . \tag{2}$$

Taking the mean and variance of M consecutive traffic envelopes allows the variability of the traffic envelope itself to be characterised at longer time-scales.

From the traffic envelope, this MBE approach computes two estimates of effective bandwidth E, one for each of the two time-scales: short-term burstiness and long-term variance. For the long-term time-scale resulting from variance between traffic envelopes, the mean and standard deviation of the maximal traffic envelopes (those measured over $T \cdot \tau$) provide one estimate of the effective bandwidth,

$$E_{\text{long}} = \overline{R_T} + \alpha_{\text{long}} \sigma_T . \tag{3}$$

The value of α_{long} will determine how the estimator behaves in response to variability in the measured flow. It is possible to formulate α_{long} to dictate a specific confidence interval for these constraints. [10] considered a large variety of distributions on which to base α_{long} -- settling upon a Gumbel distribution for its ability to describe the asymptotes of the extremes for a large range of other distributions (e.g. Gaussian, exponential, log-normal, Gamma, Raleigh). However other work -- [25] and [26] -- indicated that a Gaussian distribution is adequate, as well as allowing a more tractable computation. Thus in each case the computation of α_{long} is based upon computing the inverse of a complementary CDF of an N(0,1) Gaussian distribution $(Q^{-1}(\cdot))$ based upon the maximum packet loss (ε) and the traffic envelope:

$$\alpha_{\text{long}} = Q^{-1}\left(\frac{\varepsilon \overline{R_T}}{\sigma_T}\right) . \tag{4}$$

For the shorter burstiness time-scale, a different estimator is is used. The effective bandwidth requirement of the burst time-scale relates to the size of the buffer, q. The estimate of effective bandwidth requirement is computed from the maximum of the traffic envelope mean and standard deviation. In the following equation, C -- the capacity of the link -- is required to compute the rate at which the buffer can be drained:

$$E_{\text{short}} = \max_{k=1,2,\ldots,T}\{\frac{(\overline{R_k} + \alpha_{\text{short}}\sigma_k)kT}{k\tau - \frac{q}{C}}\}. \quad (5)$$

Unlike E_{long}, E_{short} is computed using every value of k in the traffic envelope. Once again, the standard deviation pre-multiplier will determine the response to variability in the measured flow. The derivation of α_{short} from the user supplied packet-loss, ε, and traffic envelope is

$$\alpha_{\text{short}} = Q^{-1}(\frac{\varepsilon\overline{R_T}}{\sigma_k}). \quad (6)$$

The maximum of the two equations 3 and 5 can be considered the worst-case effective bandwidth estimate of the traffic flow described by the traffic envelope. This is given by

$$E = \max\{E_{\text{long}}, E_{\text{short}}\}. \quad (7)$$

[10] documents the importance of the value of T, the maximum number of samples for a traffic envelope. An ideal value of T will provide the optimum use of resources, while too small a value of T causes the variation over σ_T to be large, so that the capacity-based estimate of Equation 3 will be pessimistic. Alternatively, if T is too big the estimate derived for buffer occupancy will be too large causing the buffer based estimate, Equation 5, to be pessimistic. In [10] a discussion is given over to locating the optimum value of T, a value typically on the order of a few seconds.

By using the ability to nominate queue size and overflow probability, service allocations can be computed for certain queue sizes. The boundary on the delay experienced through the buffering of any packet in a flow may be considered as the transmission time per packet multiplied by the capacity of the queue. As a result the ability to compute maximum buffer sizes from delay constraints allows the computation of service allocations treating the overflow probability as the same probability that packets will be delayed beyond the delay-bound.

4 Implementation

The scheduler implements a guaranteed fair-service queueing algorithm to bound queueing delays. The WF^{2}Q+ supplies a weighted service for queued traffic with weights corresponding to the amount of service (link bandwidth) each aggregate-flow of traffic may use. The facilities of this scheduling algorithm mean that, for the implementation, no regard need be given to the potential delay of large weight values. Additionally, because the scheduler is work conserving for traffic that is not delay bound, there is no wasted resource: the scheduler will where appropriate reallocate any unused resource among queues with packets requiring service.

Traffic flowing through the network-element is measured as inputs for an MBE. Using allocation-policy nominated control parameters, either target loss-ratio or delay-bounds, the MBE computes resource requirements for each class of traffic. The available resource is then divided up, using a weighted value derived from these estimates, and each appropriate weighted value is then installed into the network-element's scheduler. This process is continuously repeated, updating the weights of the WF^{2}Q+ values dynamically, as the traffic characteristics change.

ı addition to allocating scheduler resource, it is possible to compute a queue ize for a given loss rate and link capacity combination, as would be the case ɔr a traffic-class with a guaranteed throughput. The computation of the apacity is given as

$$\overline{}_T + \alpha_{\text{long}}\sigma_T = C, \tag{8}$$

ıere α_{long} is given in Equation 4. While an estimate of the queue size q is iven by:

$$\max_{=1,2,\ldots,T}\{k\tau(\overline{R_k} + \alpha_{\text{short}}\sigma_k - C)\} = q, \tag{9}$$

here Equation 6 defines the value of α_{short}. For this implementation, previous xperience with active buffer management in partially-shared buffers, [27,28], ndicated that to ensure that there is an adequate differentiation between raffic, such systems are sensitive to the load of each traffic type in the uffer and to the actual threshold value used.

s a result, the approach taken here is different. The buffer sizing is not used s a principal mechanism to differentiate one session from another. Instead, uffer sizing is used principally as an upper-bound on the delay properties of raffic where appropriate. If traffic is delay sensitive it stands to reason hat traffic delayed by more than a nominated amount has no value and that raffic exceeding this delay ought to be discarded. In contrast, the traffic may ot be discarded if it exceeds the buffer thresholding values for flows that do ot have an explicit delay constraint. This approach makes available ransmission capacity that would have otherwise been wasted on traffic that was utside its delay constraint.

his scheme may be thought of as a form of work-conservation for the flows that ave no delay-constraint but non-work-conservation for flows that do have a elay-constraint. The link-capacity that may be wasted on the delay constrained raffic with packets now too delayed to be of use are used by the traffic that as no such delay-constraint.

Experimental Method

his section details the environment used for the experiments of Section 6. The ardware implementation was based around a commercially available network lement.

he test-environment consists of a combination of hardware and software. The ardware consists of the network-element (switch) and network interface cards. he software was written to obtain measurements from the network element, ompute new configurations of flow-weights and buffer depths, generate network raffic and control the generation of traffic sources. Figure 2 shows the mplementation architecture adopted to evaluate the dynamic allocator scheme.

igure 2 illustrates that the MBE passes estimates to the dynamic allocator, ased upon measurements of current utilisation. The allocator regularly ecomputes and updates the configuration of the network-element installing the atest configuration for scheduler weights and buffer limits.

~~n this test environment, it is possible to start flows of traffic originating~~ rom model sources, video stream sources, pre-recorded traffic flows and actual lastic traffic such as TCP/IP. Such flows are initiated and terminated without ny direct interaction with the dynamic allocator.

ıe test environment is based upon a Fore Systems ASX-200WG ATM switch as its ɛtwork-element. The traffic generators are based upon Unix workstations (for :P traffic), network-based video cameras and generators capable of creating ɱthetic workload. Computation of the estimates, as well as control of the test ıvironment, is performed by task-dedicated Unix computers. Interaction with the ɛtwork-element is done through a devolved control architecture based upon the ɔrk of [29] using extensions to the Python programming language. The components : the test environment are described more fully in [21].

:awn from the two examples of DiffServ given in Section 2, two configurations :e used to illustrate the behaviour of the MBE-based dynamic allocator. The irst configuration is based upon Olympic differentiated service using three lasses each receiving a proportion of the available link capacity. The second ɔnfiguration is based upon absolute differentiated service where three ifferent classes (one delay-bound, one loss-bound and one Best-Effort) share ɾailable resource.

ıe precise configuration of the policy is given alongside each set of results. ıe network is a dumbbell configuration with a single constriction point at the ɛtwork-element. The link capacity is configured for 100 Mbps.

Results

ıe results included in this section illustrate that the dynamic allocator is ɔle to provide a number of differentiated services across a range of guarantees ithout the need for static allocation policy. This system is able to use the ɛsources of link-capacity and buffer-space to provide service to all competing ɹality assurances with reduced resource waste. Importantly, this system ɛrforms better than best-effort by supplying differentiation. The nplementation performs better than fixed resource policy by adapting to ıanging requirements and, by adapting to changing demands, dynamic allocation ɔes not waste resources in the manner that fixed resourcing policy does.

ʌo distinct experiments are reported here, firstly Section 6.1 reports results llustrating the operation of an allocator with four classes of traffic as part f an Olympic service. The three Olympic services (Gold, Silver & Bronze) and a ɛst-effort each receive a simple priority based allocation scheme.

n contrast, Section 6.2 presents results for a set of experiments that provide uantitative assessment of the performance of the dynamic allocation mechanism, ɔr a group of traffic classes with orthogonal requirements. Using a combination f low-latency voice traffic, low-loss video traffic and a best-effort class for ɛb traffic, the flexibility and successful operation of the dynamic allocator s demonstrated.

.1 Olympic Service Operation

n this section, figures illustrating the operation of the dynamic allocator are hown. The policy used is the Olympic service detailed in Section 2.1.

he dynamic allocator in operation is illustrated in Figure 3. The top graph hows the current resource demand of the three provisioned services. The middle raph shows the allocation of the scheduler to each traffic class. Each vertical ine, representing the allocation in any particular allocation period, is ivided into up to four segments. Each segment represents the allocation of andwidth to one traffic class. The throughput experienced by each traffic class s plotted in the bottom of the three graphs of Figure 3.

n Figure 3 it is clear that at 200 seconds, an increase in the requirements for he Gold service have (virtually) eliminated any resource for a best-effort

-vice. At the 300 second mark, the resource requirements of the Silver class -e increased resulting in the Bronze service being penalised. Following a :toration in requirements of both Gold and Silver services to their former -els, service capacity is automatically made available to the Bronze service [remainder is available for the fourth, best-effort, service. It is quite ›arent that any commitments made to the Bronze service were not sustained :ween 300 and 400 seconds, although such drop-outs in service may be part of : Service Level Agreement made between the network-provider and network-users.

ıy alternatives in policy are possible. In this example strict allocation .ority is maintained. Another network-provider may implement a restriction on : impact each service may have upon another. Because the allocation system is ›grammable, as indicated in Section 2.3, the process may incorporate any ›cedure the policy dictates.

is illustrated by this example the scheme operates as required. The next :tion details the performance gained for experiments run over longer periods time. These results, made with a system offering orthogonal services, are ıpared with the performance gained using non-dynamic allocation such as best-:ort and fixed-allocation resourcing.

:—~~Orthogonal Service Performance~~

a second experiment the dynamic allocator is configured with three classes of ıffic. These classes include traffic that is delay-bound and loss-bound along :h a best-effort class intended to use the remaining available capacity. The ıamic allocator was configured to reassess the current allocation every 100 The configuration of the estimator had measurements made every 1.3 ms, with : MBE configured so that the measurements covered a period sufficiently large sample beyond the reallocation period, (for the MBE of this algorithm, $\tau =$ 3 ms, T = 200, and M = 4), therefore providing maximum protection from traffic ıctuations between consecutive allocations. The values were selected to place ıctical demands on memory, CPU and measurement systems. It is expected that :se values would have a traffic-related optimum however, it was critical to .ustrate that values dictated by the implementation environment could provide :quate results.

›le 1 lists each traffic class. Alongside the characteristics of each traffic ıss are listed the policy characteristics.

_combination of a low-delay voice aggregate, with a high-demand video jregate, consumes the majority of the available capacity. The voice traffic :rates as a continual flow arrival and departure process affording the test ıffic the full dynamics of a multiplex of voice data flow characteristics. :h voice traffic flow, VP64S23, represents a silence-suppressed voice channel :h a 64kbps peak, 23kbps mean and a mean-burst length of approximately 23068 :ets or about 60 packets (1325 octets in length). These values are derived ›m ON and OFF times of 352 ms and 650 ms from [30]. The voice-traffic class :ries a multiplex of VP64S23 flows. All active VP64S23 flows are multiplexed jether to form the traffic aggregate carried in as the first traffic class.

raffic	Flow Parameters	Policy
P64S23	300 second mean hold time, log-normal distribution; 2 flows per second mean, exponentially distributed	Delay Sensitive 1×10^{-5} ratio for packets delayed by > 500 μs (1st priority)
P25S4	4 (continuous) flows	Loss Sensitive Target loss ratio 1×10^{-4} (2nd priority)
P10S1	10 (continuous) flows	Best-effort

able 1 Parameters for traffic and policies of long duration dynamic allocator xperiments

he video data consists of four permanent streams of VP25S4. The VP25S4 video raffic is based upon an MPEG encoded, non-adaptive, video stream. Each VP25S4 as a peak rate of 25 Mbps and a sustained rate of 4 Mbps. Starting at random uncorrelated) locations in the video-stream, the multiplex of four streams of raffic provides the characteristics of high-capacity, high-throughput users ombined with the statistical effects evident both in individual traffic streams nd evident in a multiplex of strongly structured data.

he third traffic is WP10S1. This traffic consisting of TCP/IP streams, epresents an aggregate of WWW transactions, and is transmitted as the 3rd class onsuming remaining capacity. This class is elastic, using the remaining, unused apacity and as a result is affected by the ongoing availability of capacity. his source has previously been considered a multi-stage Markov chain. The ational of this approach along with the configuration parameters are drawn from he work presented in [31].

or this traffic type, performance of the available capacity can be measured by he rate at which bytes of data are able to be transferred between the elastic raffic's server and client; this performance figure is given as the goodput in he results of Table 2.

raffic	Results Mean Utilisation	Mean Allocation	Performance
est Effort (no service differentiation)			
P64S23	13.8 Mbps	---	9.4×10^{-4} packets-delayed
P25S4	17.6 Mbps	---	2.7×10^{-3} packets-lost
P10S1	10.0 Mbps		4.4 Mbps goodput
ixed Service Allocation			
P64S23	13.7 Mbps	39.3 Mbps	0 packets-delayed
P25S4	17.7 Mbps	60.7 Mbps	6.5×10^{-4} packets-lost
P10S1	0 Mbps	0 Mbps	0 Mbps goodput
ynamic Allocator			
P64S23	13.8 Mbps	27.6 Mbps	2.2×10^{-5} packets-delayed
P25S4	17.6 Mbps	71.2 Mbps	1.7×10^{-5} packets-lost
P10S1	0.8 Mbps	1.2 Mbps	314 kbps goodput

able 2 Results of long duration dynamic allocator experiments

side from the goodput, Table 2 presents the achieved loss ratio for the video tream traffic and the ratio of voice packets delayed beyond the nominated delay onstraint. This table presents results gained using best-effort, fixed llocation, and the dynamic allocator described as follows. The best-effort esults were for a system that offered no service-differentiation between the

hree different classes. Clearly, the WWW traffic WP10S1 gained excellent oodput at the expense of the loss and delay of the video and audio traffic.

'he fixed service allocation results gave the voice (in this example the highest riority) all the bandwidth required. The fixed bandwidth allocation was based pon the peak-rate requirements of the voice traffic. As VP64S23 flows start and top, the allocated resource was adapted as required. An attempt was made to llocate to the video traffic a fixed bandwidth allocation based upon its peak-ate requirements, although this was never able to be satisfied. The immediate esult for the video traffic was that there was insufficient bandwidth for an llocation that would satisfy the requirements outlined in Table 1. Finally, ith allocations made for voice and video, there was no bandwidth remaining for static allocation to the WP10S1 traffic and as a result no throughput or oodput was achieved.

'inally, the dynamic allocator results indicate that it was able to achieve the olicy agreements in this system. However, along with the mean allocation equirements, the delay/loss ratios for both the VP64S23 and VP25S4 are taken rom results taken over long-running experiments and results measured on a maller timescale may indicate lower performance. Additionally, for the dynamic llocator results, the error margin on the packet-delay figures as they were athered is still quite high, at $\pm 12\%$ with a 95% confidence interval for 1×10^{-5} nd $\pm 5\%$ with a 95% confidence interval for the packets-loss figure. xperiments for these results were run for a sufficient time to reduce the error lue to sampling to less than $\pm 1\%$ with a 95% confidence interval. Thus, taking nto account the precision of the results gained, it may be concluded that the lynamic allocator prototype worked successfully.

' Conclusions

urrently, network-elements have employed inflexible fixed resource partitioning etween different classes of traffic. Changes in network requirements, and in he traffic carried, motivated the construction of the dynamic allocator scheme resented in this description. Measurement-Based Estimation is used as input to dynamic resource allocator. Such an allocator can offer differentiated service y adjusting the service-weighting on a queue scheduler and controlling the ptimum buffer depth for queueing packets from each class. As a result, the pecification of policy allows for a highly flexible scheme. The results section llustrates the scheme in operation within an actual implementation.

'he results indicate that while this is not a general answer to all lifferentiated service network problems, this scheme provides a novel and unique pproach to offer diverse, and orthogonal services to a wide variety of traffic ypes.

?roblems may still remain with over-allocation due to delay constraints. lowever, without the support of a more complex scheduling algorithm, [32,3] the pproach taken gave acceptable results. Additionally, providing the best service or elastic traffic still requires further work.

[n spite of the TCP-related drawbacks, the dynamic allocator is able to provide service that is better than the fixed allocation approach, providing both oice and video data with the desired conditions of loss and delay. dditionally, by using the dynamic allocator in place of a fixed allocation, rovision was available for a third, best-effort service that used the left-over andwidth, a service not provided for at all in the fixed allocation approach.

References

[1] D. D. Clark, S. Shenker, L. Zhang, Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism, ACM Computer Communication Review 22 (4).

[2] S. Floyd, V. Jacobson, Link-sharing and resource management models for packet networks, IEEE/ACM Transactions on Networking 3 (4) (1995) 365-386.

[3] I. Stoica, H. Zhang, T. Ng, A hierarchical fair service curve algorithm for linksharing, in: Proceedings of ACM SIGCOMM'97, Cannes, France, 1997.

[4] V. Paxson, S. Floyd, Wide-Area Traffic: The Failure of Poisson Modeling, IEEE/ACM Transactions on Networking 3 (3) (1995) 226-244.

[5] M. E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: Evidence and possible causes, IEEE/ACM Transactions on Networking 5 (6) (1997) 835-846.

[6] A. Feldmann, A. C. Gilbert, W. Willinger, T. G. Kurtz, The Changing Nature of Network Traffic: Scaling Phenomena, ACM Computer Communication Review 28 (5) (1998) 5-29.

[7] R. J. Gibbens, F. P. Kelly, P. B. Key, A decision-theoretic approach to call admission control in ATM networks, IEEE Journal on Selected Areas in Communications 13 (6) (1995) 1101-1114.

[8] R. J. Gibbens, F. P. Kelly, Measurement-based connection admission control, in: Proceedings of 15th International Teletraffic Congress (ITC15), 1997.

[9] S. Jamin, S. J. Shenker, P. B. Danzig, Comparison of measurement-based admission control algorithms for controlled-load service, in: Proceedings of IEEE INFOCOM'97, Kobe, Japan, 1997.

[10] E. W. Knightly, J. Qiu, Measurement-based admission control with aggregate traffic envelopes, in: Proceedings of 10th Tyrrhenian International Workshop on Digital Communications, Ischia, Italy, 1998.

[11] M. Grossglauser, D. Tse, A Framework for Robust Measurement-Based Admission Control, ACM Computer Communication Review 27 (4) (1997) 237- 248.

[13] J. M. Hyman, A. A. Lazar, G. Pacifici, Real-time scheduling with quality of service constraints, IEEE Journal on Selected Areas in Communications 9 (7) (1991) 1052-1063.

[14] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification, RFC 2205, IETF (Sep. 1997).

[15] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An Architecture for Differentiated Service, RFC 2475, IETF (Dec. 1998).

[16] K. Nichols, S. Blake, F. Baker, D. Black, Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, RFC 2474, IETF (Dec. 1998).

[17] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, Assured Forwarding PHB Group, RFC 2597, IETF (Jun. 1999).

[18] V. Jacobson, K. Nichols, K. Poduri, An Expedited Forwarding PHB, RFC 2598, IETF (Jun. 1999).

[19] A. K. Parekh, R. G. Gallager, A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case, IEEE/ACM Transactions on Networking 1 (3) (1993) 344-357.

[20] D. C. Stephens, J. C. R. Bennett, H. Zhang, Implementing scheduling algorithms in high-speed networks, IEEE Journal on Selected Areas in Communications 17 (6) (1999) 1145-1158.

[21] A. Moore, S. Crosby, An experimental configuration for the evaluation of CAC algorithms, Performance Evaluation Review 27 (3) (1999) 43-54.

[22] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell, F. Toomey, Entropy of ATM traffic streams, IEEE Journal on Selected Areas in Communications 13 (6) (1995) 981-990.

[24] J. Qiu, E. W. Knightly, QoS control via robust envelope-based MBAC, in: Proceedings of 7th IEEE/IFIP Workshop on Quality of Service IWQoS, Napa, CA, 1998.

[25] J. Qiu, Measurement-Based Admission Control in Integrated-Services ~~Networks, Master's thesis, Electrical and Computer Engineering, Rice University,~~ Houston, Texas (Apr. 1998).

[26] D. Tse, M. Grossglauser, A Time-Scale Decomposition Approach to Measurement-Based Admission Control, in: Proceedings of IEEE INFOCOM'99, New York, NY, 1999.

[27] H. Kroner, G. Hebuterne, P. Boyer, A. Gravey, Priority Management in ATM Switching Nodes, IEEE Journal on Selected Areas in Communications 9 (3) (1991) 418-427.

[28] N. Matsufuru, R. Aibara, Flexible QoS control using partial buffer sharing with UPC, IEICE Transactions on Communications E83-B (2) (2000) 196-203.

[29] S. Rooney, J. E. van der Merwe, S. Crosby, I. Leslie, The Tempest, a Framework for Safe, Resource Assured, Programmable Networks, IEEE Communications Magazine 36 (10) (1998) 42-53.

[30] P. T. Brady, A model for generating ON-OFF speech patterns in two-way conversations, The Bell System Technical Journal 48 (9) (1969) 2445-2472.

[31] P. Barford, M. Crovella, Generating Representative Web Workloads for Network and Server Performance Evaluation, in: Proceedings of ACM SIGMETRICS'98, Madison, WI, 1998.

[32] H. Sariowan, R. L. Cruz, G. C. Polyzos, Scheduling for quality of service guarantees via service curves, in: Proceedings of the International Conference on Computer Communications and Networks (ICCCN), Las Vegas, NV, 1995.

Claims:

1. An apparatus for providing communications network resource to a plurality of classes of use of the network, a different level of service being associated with each said class of use, said apparatus comprising: means for estimating the demand for each of said plurality of classes of use; a dynamic resource allocator for allocating to each class a proportion of said communications network resource, the proportion allocated being dependent on the estimated demand for each class, the allocation optimising use of the available resource whilst at the same time ensuring that the level of service of each class is observed; and a communications network element for providing to each class the proportion of network resource allocated to it.

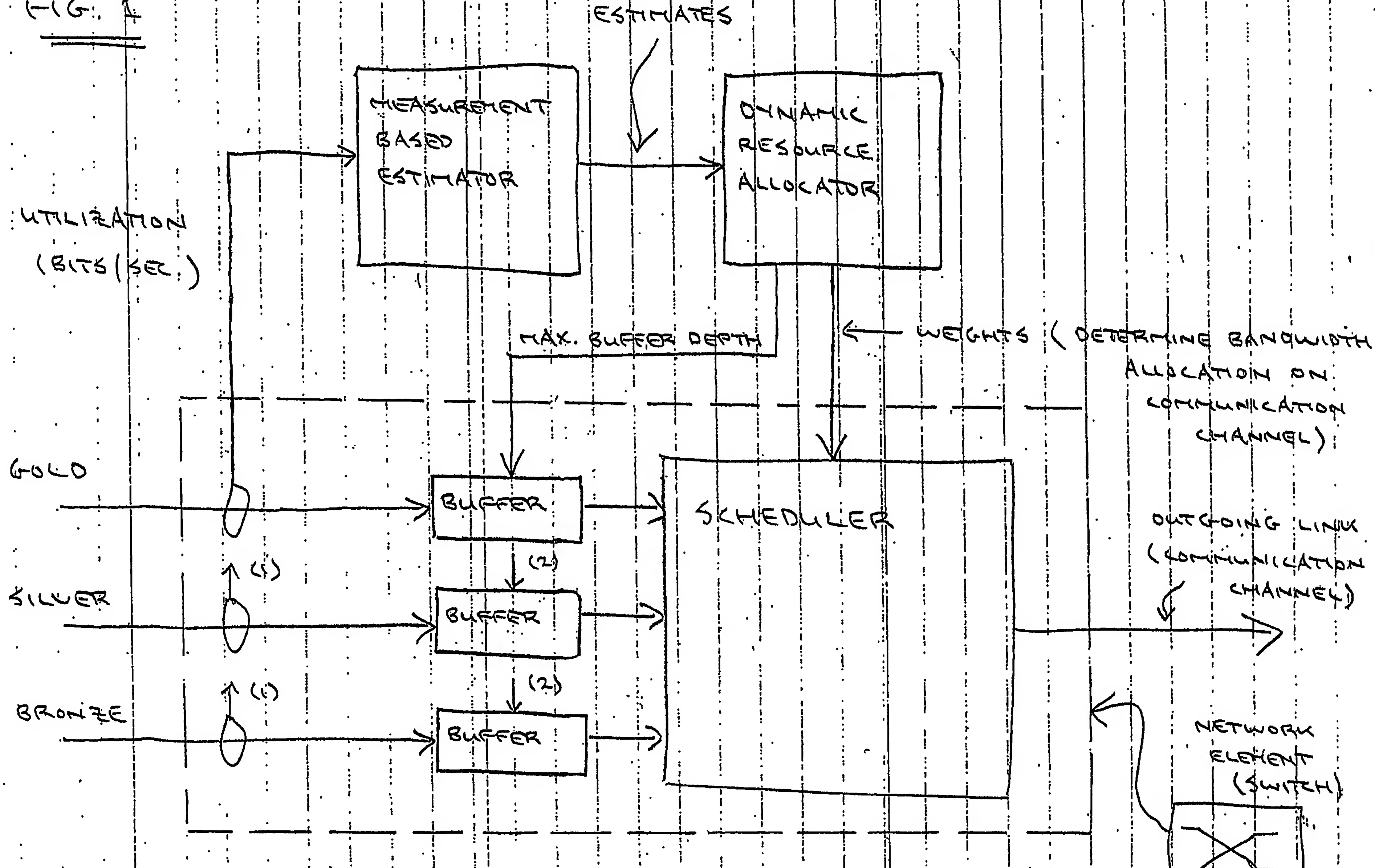
2. An apparatus according to claim 1 wherein said communications network resource comprises bandwidth of a communications channel fed by said network element and/or buffer depth in said network element.

Abstract

An apparatus for providing communications network resource

Dynamic allocation of network resource through the use of a measurement-based estimator is described. Measurements of bandwidth utilisation allow a measurement-based estimator to compute the bandwidth requirements of the measured traffic. The use of such an estimator allows provision of differentiated services by adjusting the service-weighting of a queue scheduler and modify the depth and behaviour of buffering. By providing a dynamic allocation of resource, the technique makes possible the differentiation of diverse traffic types with a reduction in the complexity and waste of current techniques such as static-allocation or the best-effort service common in the Internet. A novel approach is described to problems arising from the desire to offer diverse and sometimes orthogonal service facilities to a wide variety of traffic types.

FIG. 1



(1) TO MEASUREMENT BASED ESTIMATOR (UTILIZATION)

(2) FROM DYNAMIC RESOURCE ALLOCATOR (MAX. BUFFER DEPTH)

Figure 2: Test environment

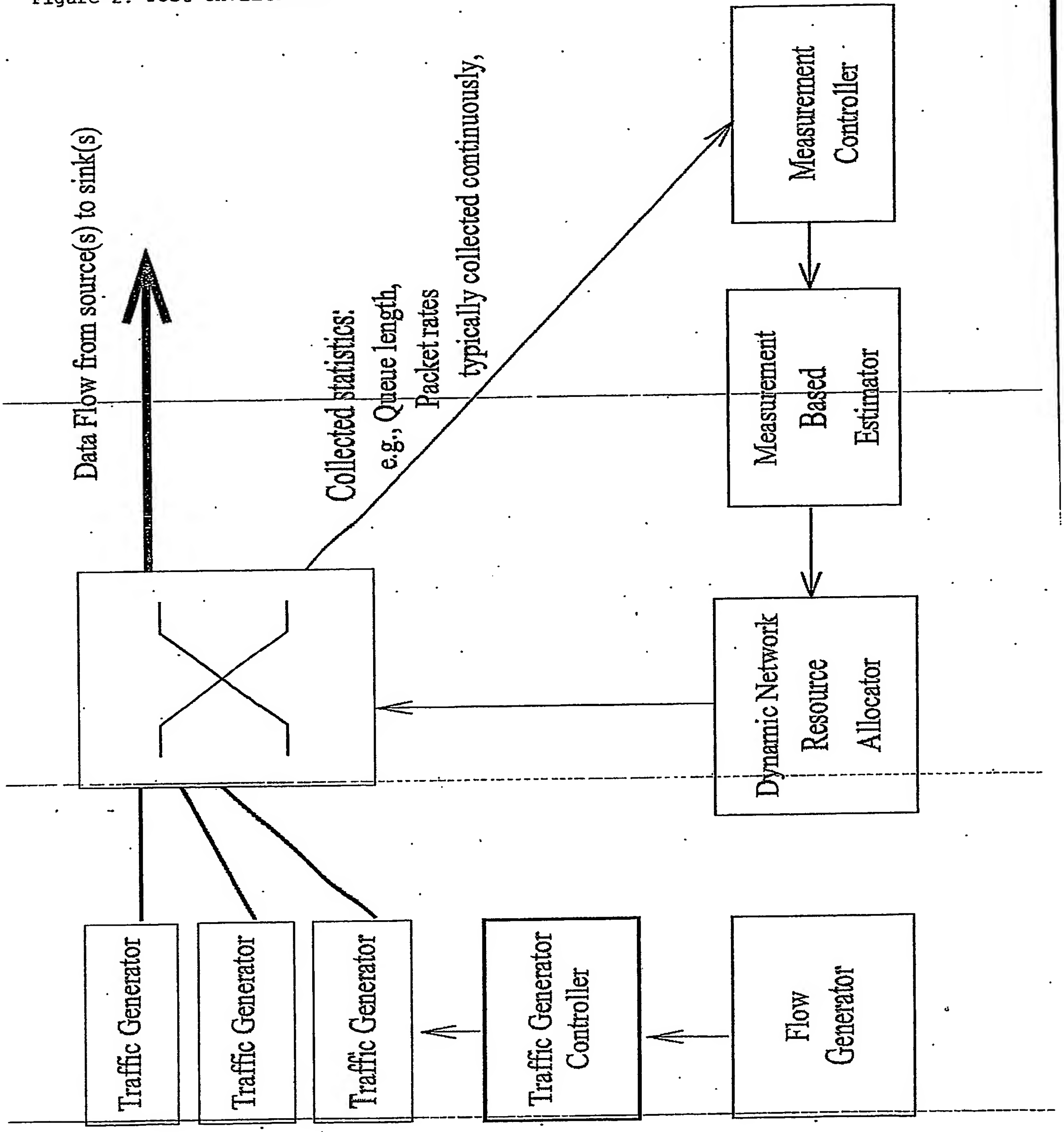


Figure 3: Olympic service